PAC Quasi-automatizability of Resolution over Restricted Distributions

Brendan Juba*
Harvard University
bjuba@alum.mit.edu

April 18, 2013

Abstract

We consider principled alternatives to unsupervised learning in data mining by situating the learning task in the context of the subsequent analysis task. Specifically, we consider a query-answering (hypothesis-testing) task: In the combined task, we decide whether an input query formula is satisfied over a background distribution by using input examples directly, rather than invoking a two-stage process in which *(i)* rules over the distribution are learned by an unsupervised learning algorithm and *(ii)* a reasoning algorithm decides whether or not the query formula follows from the learned rules. In a previous work [15], we observed that the learning task could satisfy numerous desirable criteria in this combined context – effectively matching what could be achieved by agnostic learning of CNFs from partial information – that are not known to be achievable directly. In this work, we show that likewise, there are reasoning tasks that are achievable in such a combined context that are not known to be achievable directly (and indeed, have been seriously conjectured to be impossible, cf. Alekhnovich and Razborov [1]). Namely, we test for a resolution proof of the query formula of a given size in quasipolynomial time (that is, *"quasi-automatizing"* resolution). The learning setting we consider is a partial-information, restricted-distribution setting that generalizes learning parities over the uniform distribution from partial information, another task that is known not to be achievable directly in various models (cf. Ben-David and Dichterman [5] and Michael [20]).

*Supported by ONR grant number N000141210358.

1 Introduction

When learning is employed in data mining and artificial intelligence, the objective is not simply to form hypotheses that are supported by the data, but to draw further inferences based on these learned hypotheses. The use of logical inference on such learned hypotheses is problematic since learning algorithms (namely, PAC-learning algorithms here [25]) do not guarantee that the hypotheses they produce can be interpreted as "valid" in the standard sense of mathematical logic. Motivated by this lacuna, Valiant [26] introduced *PAC-Semantics* (for a logic) to enable the analysis of logical inferences drawn from learned hypotheses. Such logical inferences are only interesting in a partial-information model, since otherwise a query about a candidate conclusion can be evaluated directly by evaluating the query formula on each complete example and considering the fraction of examples that satisfy the query.

Valiant's work analyzed a two-stage process in which in the first stage, learning algorithms are used to extract explicit premises from examples drawn from a background distribution. ("Supervision" is only provided in that the rules predict individual bits of the examples based on others.) In the second stage, conclusions are drawn from these premises by applying logical inferences. A stronger approach would be to ask, given a query formula as input, if there exist some premises that can be verified to hold on the (incomplete) examples, for which logical inference could derive the given query. The work of Khardon and Roth [17] on *learning to reason* suggests that it may be advantageous to take the two tasks together rather than in two separate stages as done by Valiant: they show that by processing the data directly, it is possible to answer queries (in a slightly different model) for which the reasoning problem is NP-hard, and for which the hypotheses that would support such queries are general DNFs, and hence not known to be learnable. In a follow-up work, Khardon and Roth [18] showed similar results for a answering a more limited class of queries from incomplete information.

One of Khardon and Roth's major objectives in their works was to provide tractable reasoning by avoiding the use of theorem-proving techniques. Indeed, in most approaches to these problems of learning and reasoning that are actually used in practice (e.g., based on graphical models) the resulting reasoning problem is known to be #P-hard, even for highly restricted classes of formulas (cf. Roth [23]). Khardon and Roth's approach [18] could still only answer CNF queries under partial information that were, for example, Horn clauses or of low width. Our recent work [15] established that the combined problem of evaluating a query formula from partial examples using theorem-proving is actually no more difficult than the "second stage" problem of exact, classical reasoning for all "natural" (in the sense of Beame et al. [4]) fragments of proof systems.[1] Such results raised the possibility that in some cases, such theorem-proving under PAC-Semantics may actually be easier than in the classical case. The results of the current work suggest that this is actually so: we quasi-automatize[2] the (*general, dag-like*) resolution proof system under PAC-Semantics (i.e., when given incomplete examples) with restricted distributions, whereas the best known automatization of resolution, first obtained by Clegg et al. [10], achieves a running time of $n^{O(\sqrt{n \log n})}$ for polynomial-size resolution proofs over n variables. Moreover, results by Alekhnovich and Razborov [1] suggest that perhaps resolution is not quasi-automatizable at all.

[1] We note some similarly motivating results of Alekhnovich et al. [2], showing how theorem-proving algorithms can solve classical learning problems, and thus that learning is easier than theorem-proving over the same representations.

[2] *Automatization* is the standard terminology for the problem of deciding whether or not a proof of a formula exists in a given (fragment of a) proof system. *Quasi*-automatization means that there is a quasipolynomial time algorithm for this problem.

In this work (in contrast to [15]), we consider problems in which the background distribution (and hence learning problem) is restricted. The main example we consider is a generalization of learning parities under the uniform distribution: we note that one way of viewing the (standard) problem of learning parities over n bits is that there is an unknown parity constraint over the $n+1$ bits; here, we consider an "unsupervised" setting where there is again no distinguished label bit, and there may be several such constraints. We refer to such distributions as *affine distributions.* Such problems may easily be impossible under partial information: in the (weakly) Restricted Focus of Attention (wRFA) model, Ben-David and Dichterman [5] showed that learning a large parity becomes impossible if an insufficient number of bits can be simultaneously viewed. We remark that such parity formulas are also natural and interesting from the standpoint of the reasoning problem, since they presented classic examples of theorems that were hard for resolution [24].

Our theorems are actually established for a more general class of distributions that merely feature a *"correlation gap"*: when we condition on a conjunction, the other bits are either highly biased or feature bounded bias (with the bounds on these biases parameterizing the "gap" that gives these distributions their name). Such distributions are reasonably natural. For example, some simple topic models, e.g., along the lines of the *pure document* model underlying the analysis of Latent Semantic Indexing given by Papadimitriou et al. [22], feature such correlation gaps. Roughly, the presence or absence of "primary terms" in a document determine the relative prevalence or absence of other primary terms by determining the latent topic variable, and other words do not have such an effect.

1.1 A closer look at our results

To get a sense of how automatization under PAC-Semantics could possibly be easier than the usual, "worst-case" automatization, we first consider the case of automatizing over the uniform distribution in Section 3.1. We observe that clauses of at least logarithmic width have a satisfied literal – i.e., are *witnessed satisfied* – in our partial examples with high probability. Using the techniques of our previous work [15], it therefore suffices to search for proofs of small width for each of the examples, which is known to be automatizable.

We then turn to the learning context, in which there are parity constraints on the distribution that could be useful as premises in a small proof of the query. Our algorithm distinguishes when an input CNF can be refuted by a small resolution proof on the basis of some learnable formulas over the unknown distribution (e.g., some small parity constraints) from when the input CNF is satisfied with moderate probability, that is, distinguishing when its negation (a DNF) can be proved by a small resolution proof, from when its negation is falsified with moderately large probability:

Theorem 1 (Main theorem, cf. Theorem 19) *There is an algorithm that, given an input CNF φ, a bound $p(n)$, $\mu, \epsilon, \gamma, \delta \in [0,1]$, and access to partial examples from an affine distribution D in which indices are hidden independently with probability $1-\mu$ in each example, and given that either*

- *φ is satisfied by D with probability at least $\epsilon + \gamma$ or*
- *there is a resolution refutation of size $p(n)$ of $\psi_0 \wedge \psi_1 \wedge \varphi$ for CNFs ψ_0 and ψ_1 such that*
  - *ψ_0 consists of $1 - \frac{\gamma}{n^{O(\log \frac{p(n)}{\gamma})}}$-valid clauses and*
  - *ψ_1 is witnessed to evaluate to true with probability $1 - \epsilon + \gamma$ on the partial examples*

decides which case holds with probability $1-\delta$, and runs in time $n^{O(\frac{1}{\mu}\log\frac{p(n)}{\gamma\delta})}$.

Our main theorem supports the learning of two kinds of hypotheses, denoted ψ_0 and ψ_1 in the theorem statement. The conditions on ψ_0 correspond essentially to standard, "realizable" learning. Although in contrast to the usual learning set-up we allow some potentially noticeable counterexamples to the target hypothesis, these counterexamples are required to be very rare relative to the accuracy parameter γ given to the algorithm. Since the parity constraints in an affine distribution hold perfectly, the learnability of such a ψ_0 establishes the learnability of the parity constraints of an affine distribution in the context of a resolution proof, which after all can only operate on clauses. We remark that our previous work [15] could not exploit such hypotheses in general, due to the fact that we did not restrict our attention to the independent masking processes.

By contrast, the conditions required in our main theorem on the second hypothesis ψ_1 allow it to have an arbitrary error tolerance ϵ given as an input parameter to the algorithm. In this case, our algorithm is essentially detecting when there exist any CNF formulas that suffice to complete a proof of the query, that are ϵ-close to being valid. For example, in the simple topic model distributions, clauses encoding a rule that the presence of one primary term implies that at least one out of $\Omega(\log 1/\epsilon)$ of the other primary terms for the same topic also appears is a $(1-\epsilon)$-valid clause that might be useful in reasoning about such documents. The permission of an arbitrary ϵ flaw in the formulas here may be seen as analogous to the *agnostic learning* setting in traditional concept learning [16] and we feel that it is crucial to the utility of such algorithms in artificial intelligence, for example. The conditions we impose on this second hypothesis follow the conceptual developments from our previous work [15] and this groundwork merely serves as a starting point in the present setting. Still, for the reader's sake, we will briefly review why we deem the restrictions imposed on this hypothesis to be reasonable and how they lead to a tractable learning problem. Agnostic learning is notoriously hard, and the main reason we avoid this problem is by declining to specify *any* explicit representation of a relation that is ϵ-close to valid,[3] a problem that is conveniently irrelevant to our ultimate goal of answering a query here. The more serious difficulty that impacts the class of hypotheses we ultimately ask to learn is that determining whether or not a formula is satisfied in the context of partial information can be intractable or even impossible, depending on the level of generality; in our previous work we observed that this leads to a serious obstacle since the ability to detect when queries are provable implies that we can detect that the premises are satisfied. As this may not be feasible, it is therefore unreasonable to hope for an algorithm that utilizes *any* CNF that is ϵ-close to being valid as a hypothesis; what we *can* hope for instead (and is achieved by the previous work) is that we can utilize hypotheses that could be efficiently verified to be ϵ-close to being valid if only we had identified such a hypothesis. (Note that even in the usual complete information agnostic learning setting, as hard as it is, this condition is always satisfied.) The solution proposed in the previous work was to use a simple definition of *witnessed evaluation* that suffices to enable efficient certification that a formula is satisfied in a partial example. As illustrated in our warm-up, this definition moreover turns out to be strong enough to enable solutions to the combined learning and reasoning problem.

The core of the proof of the main theorem is a structural result for resolution proofs over correlation gap distributions: every formula with a polynomial-size resolution refutation (possibly using clauses that hold over the background distribution as additional premises) simplifies to a logarithmic-width refutation using logarithmic-width clauses that are learnable from the background distribution. Essentially, we only need to focus on clauses in the proof that cannot be

[3]Surely, this is a step beyond merely specifying another explicit representation that lies outside the target concept class as done in improper learning, and hence we count it as a separate issue.

learned, and the key observation is that the only way a clause of sufficient (logarithmic) width could not be witnessable (and so unlearnable) is if many of its literals are biased to be false; this may happen if the literals of the clause appear in some even-parity constraint, for example. There is then some other (small) learnable clause that can be used to eliminate each such literal, and so by a sequence of resolution steps, we can reduce the wide clauses of the proof down to logarithmic width. We similarly establish the learnability of the nearly-realizable ψ_0 by showing that we only need to consider logarithmic-width subclauses for the proof. These techniques actually also allow us to determine whether or not a given CNF query is highly valid directly: In Section 4, we exhibit an algorithm that distinguishes when an input CNF is satisfied with high probability over the background distribution from when it is falsified with some moderate probability, without reference to resolution proofs. This latter result may be understood as lifting the restrictions on the CNF query in Khardon and Roth's work [18], given some new restrictions on the distribution and obscuring of information.[4] These techniques for learning to reason about CNFs from partial information are new to this work, and seem to rely on both the correlation gap and the relative simplicity of the process that hides information.

2 Our setting: background and approach

PAC-Semantics. PAC-Semantics (for a logic) was introduced by Valiant [26] to capture the kind of validity possessed by statements produced by PAC-learning algorithms; for example, if one uses a PAC-learning algorithm to learn a conjunction $\wedge_j x_{i_j}$ to match the "labels" given by another variable x_t from examples from a distribution D, then the formula $[\wedge_j x_{i_j} \equiv x_t]$ is (probably) approximately valid in the following sense:

Definition 2 ($(1-\epsilon)$-valid) *Given a distribution D over $\{0,1\}^n$, we say that a Boolean formula φ is* $(1-\epsilon)$-valid *if $\Pr_{x\in D}[\varphi(x)=1] \geq 1-\epsilon$. If $\epsilon = 0$, we say that the formula is* perfectly valid.

Of course, the definition makes sense for other kinds of formulas (not just equivalences). It is not hard to show (by a union bound) that any classical logical inference can be applied to formulas possessing this weaker kind of validity, as long as we allow for further loss in the approximation.[5]

Proposition 3 (Classical reasoning in PAC-Semantics [15]) *Let $\psi_1, \ldots, \psi_k$ be formulas such that each ψ_i is $(1-\epsilon_i)$-valid under a common distribution D for some $\epsilon_i \in [0,1]$. Suppose that $\{\psi_1, \ldots, \psi_k\} \models \varphi$ (in the classical sense). Then φ is $1-\epsilon'$-valid under D for $\epsilon' = \sum_i \epsilon_i$.*

The main problem that we wish to address, introduced in prior work [15], is that of deciding the degree of validity of a given *query* formula using *examples* from an unknown distribution. In the present work, we will see how to obtain stronger results (than [15] in particular) for a moderately restricted class of distributions, in which the correlation between variables is either strong or weak (and not of moderate strength):

[4] By contrast, our main theorem essentially certifies the validity of DNF queries, which seems to require the existence of simple proofs in the context of incomplete information—a further discussion of this point appears in Section 2, in the context of the formal set-up.

[5] It also is not hard to show that as long as the distributions are arbitrary, the union bound is tight here [15].

Definition 4 (Correlation gap) *We will say that a distribution D over $\{0,1\}^n$ has a* width-w $(\beta, 1-\gamma)$ correlation gap *if for any conjunction of $k \leq w$ literals $\ell_1 \wedge \ldots \wedge \ell_k$ satisfied with nonzero probability and any variable x, either $\Pr[x = 1 | \ell_1 \wedge \cdots \wedge \ell_k] \geq \beta$ and $\Pr[x = 0 | \ell_1 \wedge \cdots \wedge \ell_k] \geq \beta$ or else for some $b \in \{0,1\}$, $\Pr[x = b | \ell_1 \wedge \cdots \wedge \ell_k] \geq 1 - \gamma$. In the former case, we say that x is β-*balanced for $\ell_1 \wedge \cdots \wedge \ell_k$, *and in the latter we say that, respectively, x (for $b = 1$) or $\neg x$ (for $b = 0$) is $1 - \gamma$-*implied by $\ell_1 \wedge \cdots \wedge \ell_k$, *denoted as $\ell_1 \wedge \cdots \wedge \ell_k \rightsquigarrow x$ (or $\ell_1 \wedge \cdots \wedge \ell_k \rightsquigarrow \neg x$, respectively).*

A simple example of a large collection of distributions that have a very strong correlation gap of $(1/2, 1)$ up to width n and serve as a motivating example, is the class of *affine distributions*:

Definition 5 (Affine distribution) *For any solvable linear system over $\mathbb{F}_2$ $Ax = b$, the distribution over $\{0,1\}^n$ that is uniform over solutions to the linear system is an* affine distribution.

Affine distributions turn out to have a correlation gap, since intuitively, if a conjunction determines the value of another variable in a linear constraint, that literal is easily seen to be 1-implied, and otherwise the literal turns out to be uniformly distributed (and thus, 1/2-balanced).

Definition 6 (Constraints on clauses) *We say that there is a* constraint *on a clause C in an affine distribution given by the linear system $Ax = b$ if there is a linear combination of the rows of A such that the only nonzero entries are in indices i for which the corresponding variable appears in a literal of C.*

Lemma 7 *Let C be a clause and D be an affine distribution such that there are no constraints on C. Then the marginal distribution over the variables appearing in C is uniform.*

Proof: Let $Ax = b$ be the linear system defining D. Let y denote the variables in C and z denote the variables not appearing in C, and let A' be the submatrix of A formed by columns corresponding to variables in C, and A'' be the submatrix formed by columns of A corresponding to variables not in C. Note that by converting the matrix into reduced form, we may verify that if for some y^* there were no solutions to the system $A''z = (b + A'y^*)$, then there must be some linear combination of the rows of A'' yielding the 0 vector. The corresponding linear combination of the rows of the linear system $Ax = b$ would be a constraint on C, which does not exist by hypothesis. Therefore, for every y^*, the system $A''z = (b + A'y^*)$ has solutions, and furthermore, since A'' is the same for every assignment y^*, it always has the same number of solutions. Thus, as D is uniform over these solutions (and y^*), the marginal distribution over the variables in C is indeed uniform as claimed. ■

Correlation gap distributions are reasonably natural. For example, the simple *"pure document"* probabilistic corpus model introduced by Papadimitriou et al. [22] to analyze Latent Semantic Indexing generally features a nontrivial correlation gap:

Example 8 (Pure document topic model [22]) *The* pure document topic model *is a probabilistic model of document generation. We suppose that documents are represented by the set of words appearing in them. A document is then generated in a two-stage process in which a (latent)* topic variable *is first sampled, where each outcome of the topic variable is associated with one member of a family of disjoint sets of* primary terms. *The set of words associated with this topic is obtained by the union of these primary terms with a set of generic terms (shared across topics); each word has an associated probability in the range $[\epsilon, \tau]$ for some small constants ϵ and τ. The*

document itself is then sampled by independently tossing a biased coin for each word in the topic set, including it with its given probability. If the overall probability of the topic words appearing in the document is at most $\delta \ll \epsilon$, which holds, for example, if each topic has sufficiently small probability of being chosen (relative to ϵ/τ), then the distribution has a width-w $(\epsilon, 1-\delta(1+\delta w))$-correlation gap.

Partial assignments. Answering queries using (complete) examples drawn from a distribution D is trivially accomplished. By contrast, the task is much more difficult when some of the variables' settings are deleted from the examples, resulting in *partial* example assignments. In the present work, we will focus on a very simple process for generating partial assignments that chooses whether to delete each entry from the assignment by tossing a biased coin. In learning theory, such a model first appeared in the work of Decatur and Gennaro [11]. Formally:

Definition 9 (Partial assignments and masking) *A* partial assignment ρ *is an element of* $\{0,1,*\}^n$. *We say that a partial assignment ρ is* consistent *with an assignment x if whenever $\rho_i \neq *$, $\rho_i = x_i$.*

A mask *is a function taking assignments to consistent partial assignments. A* masking process *is a mask-valued random variable. We denote the distribution obtained by applying a masking process M to a distribution over assignments D by $M(D)$.*

The independent masking process with parameter μ *is the following masking process M_μ: for every $x \in \{0,1\}^n$, for every i, $M_\mu(x)_i = x_i$ independently with probability μ, and otherwise $M_\mu(x)_i = *$.*

We note that a related set-up appears in recent works by Dvir et al. [13], Wigderson and Yehudayoff [27], and Moitra and Saks [21] with the latter result(s) being particularly relevant: in our terms, they essentially show that there is a polynomial time algorithm (in $|\mathrm{supp}(D)|$, $\frac{1}{\epsilon}, \log\frac{1}{\delta}$, and n, for constant $\mu \in (0,1]$) that for any distribution D on $\{0,1\}^n$, completely recovers D to within an additive ϵ at each point in $\mathrm{supp}(D)$ given access to $M_\mu(D)$. As a consequence, they obtain (in particular) an algorithm for answering queries that is efficient when D has sparse support: we simply recover the distribution (to within an additive $\epsilon = \gamma/|\mathrm{supp}(D)|$ error, say) and add up the mass assigned to points where the formula is true to obtain an additive γ estimate of the degree of validity of the query under D.

We will be interested in evaluating queries for distributions that have exponentially large support, such as affine distributions. In order to do so, we will need to impose further restrictions on the formula, namely (looking ahead) that the formula has a small resolution proof from clauses that are either always true or merely *witnessed* to be true on most examples from $M_\mu(D)$:

Definition 10 (Witnessing) *We define a clause to be* witnessed to evaluate to true *on a partial assignment if there is some literal $\ell(x_i)$ (x_i or $\neg x_i$) in the clause such that ($\rho_i \neq *$ and) $\ell(\rho_i) = 1$. We say that the clause is* witnessed to evaluate to false *if for every literal $\ell(x_i)$ in the clause, ($\rho_i \neq *$ and) $\ell(\rho_i) = 0$.*

A very natural and related notion is using the partial assignment as a *restriction* to simplify a formula by "plugging in the known variables." For the cases of interest (CNFs here) this operation can be simply captured by the following definition.

Definition 11 (Restricted formula) *Given a partial assignment ρ and a clause C, the* restriction of C under ρ, *denoted $C|_\rho$, is the tautological formula $\top$ if C is witnessed to evaluate to true on ρ, and otherwise it is given by the set of literals $\ell(x_i)$ in C for which $\rho_i = *$. For a CNF φ, if some clause is witnessed to evaluate to false, then the restriction of φ under ρ is defined to be the contradictory formula $\perp$; otherwise, it is the conjunction of restrictions of clauses not witnessed to evaluate to true.*

Theorem-proving versus other approaches to answering queries. When a (small) resolution proof exists, this implies that the query has the form of a DNF formula. This naturally raises the question of whether techniques similar to those used by Wigderson and Yehudayoff [27] (in particular) could be applied to the partial assignments corresponding to the DNF's terms to evaluate such a DNF directly, without turning to a resolution proof. The difficulty that arises with such an approach is that when many DNF terms correspond to partial assignments that are consistent with one another, determining the weight of their disjunction by such techniques seems to involve an (exponentially large) inclusion-exclusion expression.[6] While one could try to bound the number of terms in the sum by considering only terms of bounded width using the techniques we develop here (cf. Lemmas 15 and 17 and the approach of Theorem 19), decreasing the error per omitted term by increasing the width bound increases the number of terms at a greater rate, so that a nontrivial bound on the error cannot be obtained. By contrast, the existence of a small a priori bound on the size of a resolution proof of the formula enables us to derive a related bound on the width we need to consider (naturally, this is made precise in the proof of Theorem 19).

Khardon and Roth [18], in their work on "learning to reason," had a different approach to an essentially similar problem—among others, they considered a problem where one is given a distribution over partial assignments that are consistent with satisfying assignments of some background formula (generally a polynomial-size DNF), and they wish to decide whether the query formula is either entailed by the background formula or whether it is falsified with some given probability (promised that one of these two cases holds). When the query is a k-CNF (say for $k = O(\log n)$), they answer the query by constructing a relatively small set of partial assignments and testing whether the query has an unsatisfied clause on any of the partial assignments; this approach again answers queries while avoiding any consideration of theorem-proving. The size of the set (and therefore also the running time) is polynomial in n and exponential in k, and so it is quasipolynomial for $k = O(\log n)$. It turns out that our work can be used to extend their scheme to answer general CNF queries under distributions with strong correlation gaps and independent masking by first reducing the width of the (important) clauses in the query to $O(\log n)$, as we show in Section 4. The reader should note that this approach only works for testing whether the query is *almost perfectly* valid.[7]

We stress that these two approaches apply to largely orthogonal classes of formulas: the latter approach answers CNF queries, whereas the former, resolution-based approach answers DNF queries. For both approaches, there is an important asymmetry: in the first approach, it is the existence of a small resolution proof that restricts the class of formulas that can be used as hypotheses and queries, whereas for the second, it is a matter of the asymmetric way the error arises when testing the query on a partial assignment.

[6] Wigderson and Yehudayoff avoid the need for inclusion-exclusion by only considering complete assignments, which of course correspond to disjoint events.

[7] Although the masking process is simple enough to allow us to compute (e.g.) unbiased estimates of the probability that a given (set of) clause(s) is witnessed unsatisfied, recovering an estimate of the true probability that the formula is unsatisfied seems to again involve a potentially exponentially large inclusion-exclusion calculation.

Resolution. We now briefly review the resolution proof system, which plays a central role in this work. Resolution is a proof system for establishing the *validity* of a *DNF*: it is given by a *refutation* of its (CNF) negation.

Definition 12 (Resolution) *A* resolution refutation *of a CNF φ is given by a sequence of clauses $\{C_i\}_{i=1}^k$ such that C_k is the (unsatisfiable) empty clause $\perp$ and for each C_i in the proof, one of the following holds:*

1. (Axiom) *C_i is a clause of φ*
2. (Weakening) *some C_j for $j < i$ is a subclause of C_i*
3. (Cut) *there exist clauses $C_j = D \vee x$ and $C_\ell = E \vee \neg x$ (for some variable x) with $j, \ell < i$ such that $C_i = D \vee E$. The literals x and $\neg x$ are said to be a* complementary pair.

Note that resolution is sound: each C_i derived in a step of the proof is a consequence of $\varphi \wedge C_1 \wedge \cdots \wedge C_{i-1}$. Thus, the derivation of the empty clause implies that φ is unsatisfiable. It is standard to note that weakening is unnecessary in a resolution refutation: if we simulate a resolution proof without using the weakening rule (applying the cut rule to the original clauses instead of weakened ones at each step) we end up with a legal derivation that has a subclause of the original derivation at each step, which must also end with the empty clause. We include weakening as a derivation rule because it is useful in the analysis of resolution. In particular, we will be interested in the result of "plugging in" a partial assignment to each step of a resolution refutation:

Definition 13 (Restricted proof) *Given a resolution refutation Π and partial assignment ρ, the* restriction *of Π under ρ, denoted $\Pi|_\rho$, is the proof obtained by substituting $C|_\rho$ for each clause C appearing in Π.*

Using the weakening rule, the following (folklore) observation is easily established.

Proposition 14 (Resolution is restriction-closed) *For any CNF φ with a resolution refutation Π and any partial assignment ρ, the restriction $\Pi|_\rho$ is a resolution refutation of $\varphi|_\rho$.*

We will see the value of Proposition 14 illustrated simply in Section 3.1. (It will, of course, also play a key role in establishing the main theorem, Theorem 19.)

3 Automatizability of resolution

3.1 Width-based automatizability for the uniform distribution

We start with the simpler special case of the uniform distribution (which we denote U_n), which will introduce a useful lemma for the general case, and help develop some intuition. The key observation here is that clauses that are sufficiently wide are almost always witnessed to evaluate to true:

Lemma 15 *Let C be a clause such that for any literal ℓ of C and any subclause C' of C without ℓ, ℓ is β-balanced for $\neg C'$, and suppose C has width at least $\frac{1}{\mu\beta} \ln \frac{1}{\delta}$. Then C is witnessed to evaluate to true on $M_\mu(D)$ with probability $1 - \delta$.*

Proof: We note that ρ drawn from $M_\mu(D)$ are of the form $\rho = m(x)$ for m drawn from M_μ and x drawn from D independently. Suppose we construct a sequence of literals and subclauses of C as we sample m and x in the following way: put $C_0 = \bot$, and fix the entries of m in order until we encounter some unmasked entry corresponding to some literal ℓ_i in C; we then put ℓ_i in $C_{i+1} = C_i \vee \ell$. Each literal of C is thus included in some C_i independently with probability μ. Now, C is witnessed true on ρ precisely when some ℓ_i is set to true in x, and if the first $i-1$ literals are set to false in x, then since ℓ_i is β-balanced for $\neg C_i$, each ℓ_i is set to true in x with probability at least β when the first $i-1$ are set to false. Thus, the probability that none of these literals in a clause of width w is satisfied by ρ is at most $(1-\mu\beta)^w$, which for $w \geq \frac{1}{\mu\beta}\ln\frac{1}{\delta}$ is at most δ. ■

Since the uniform distribution is 1/2-balanced, Lemma 15 establishes that any wide clauses that appear in a resolution refutation are witnessed to evaluate to true with high probability over $M_\mu(U_n)$, and hence in a proof Π of size $p(n)$, a union bound establishes that every clause of width $\Omega(\frac{1}{\mu}\log\frac{p(n)}{\delta}))$ is substituted by $\top$ in the restriction $\Pi|_\rho$. Thus, the width-based algorithm first studied by Galil [14] that searches for refutations using clauses of width at most w (and thus runs in time $O(n^{2w})$) finds a refutation when one exists with probability $1-\delta$. The following theorem is then almost immediate:

Theorem 16 *There is a quasipolynomial time algorithm that, given an input CNF φ, size bound $p(n)$, $\mu, \epsilon, \gamma \in [0,1]$, and access to examples from $M_\mu(U_n)$, and given that either*

- *φ is satisfied by U_n with probability at least $\epsilon + \gamma$ or*
- *there exists a CNF ψ that is witnessed to evaluate to true with probability at least $1 - \epsilon + \gamma$ under $M_\mu(U_n)$ such that there is a resolution refutation of size $p(n)$ of $\psi \wedge \varphi$*

decides which case holds with probability $1-\delta$.

Proof: The algorithm takes a sample of partial assignments of size $m = \frac{1}{\gamma^2}\log\frac{1}{\delta}$ and for each partial assignment $\rho^{(i)}$ uses the width-based algorithm for $w = O(\frac{1}{\mu}\log\frac{m\cdot p(n)}{\delta})$ to check for a width-w refutation of $\varphi|_{\rho^{(i)}}$; if more than an ϵ fraction of these refutations fail, the algorithm rejects and otherwise it accepts. It is immediate that (for constant μ) the algorithm runs in quasipolynomial time in n, $1/\gamma$, $\log 1/\epsilon$, and $1/\delta$. We thus turn to considering correctness.

If φ is $\epsilon + \gamma$ valid, then Hoeffding's inequality shows that with probability $1-\delta$, at least an ϵ fraction of the assignments drawn from U_n satisfy φ; since this satisfying assignment $x^{(i)}$ is consistent with any partial assignment $\rho^{(i)}$ in the support of $M_\mu(x^{(i)})$, $\varphi|_{\rho^{(i)}}$ is satisfiable, and hence by the soundness of resolution, no refutation exists for these partial assignments, and we see that the algorithm rejects.

In the second case, we note first that (again by Hoeffding's inequality) with probability $1-\delta/2$, every clause of the unknown formula ψ is witnessed to evaluate to true in at least a $1-\epsilon$-fraction of the partial assignments. Let Π be the size $p(n)$ refutation of $\psi \wedge \varphi$. It then follows from Lemma 15 (cf. the above discussion) that with probability $1-\delta/2$, for this $1-\epsilon$ fraction of the $\rho^{(i)}$ (out of m), $\Pi|_{\rho^{(i)}}$ is a width-w refutation of $\varphi|_{\rho^{(i)}}$ since the clauses from ψ and clauses of Π of width greater than w all simplify to $\top$. Thus, in this case, the algorithm accepts with probability $1-\delta$, as needed. ■

3.2 Augmenting the width-based algorithm with learning

While the uniform distribution illustrates how the *reasoning* problem may become easier in the context of a distribution, in a sense there is no *learning* problem if the distribution is known to

be uniform: a given formula indicates which settings of the variables are "positive" or "negative" examples, and the entire learning question for a given formula merely concerns whether the distribution assigns high weight to the negative examples. We now turn to considering our learning problem.

Distributions with a correlation gap turn out to be easy to work with because we only need to consider narrow clauses: for starters, Lemma 15 guarantees that (sub)clauses of sufficient width ($\Omega(\frac{1}{\mu\beta}\log\frac{1}{\delta})$ here) for which every variable is balanced in D (for every small subset of the rest) are witnessed with high probability. Naturally, in an affine distribution, we can make a similar claim if all of the literals of a sufficiently wide clause are each involved in some constraint that 1-implies one of them. More generally:

Lemma 17 *Let C be a clause of width $2w$ for $w \geq \frac{1}{\mu}\ln\frac{1}{\delta}$ and D be a distribution with a width-w $(\beta, 1-\gamma)$ correlation gap such that for every subclause of C of size w there is some further subclause $C' \vee \ell$ with $\neg C' \rightsquigarrow \ell$. Then with probability $1-\delta$, there is an unmasked literal of C that is true with probability $(1-\gamma)$ under D.*

Proof: Let C_0 be the first w literals of C, and suppose m is drawn from M_μ and x is drawn from D. Given C_i (of width w), some ℓ_i for the subclause $C_i' \vee \ell_i$ is $(1-\gamma)$-implied by $\neg C_i'$. Thus, with probability at least $1-\gamma$, either ℓ_i is satisfied in x or some other literal of C_i' is satisfied in x. We let ℓ_i' be the other satisfied literal of C_i' in the latter case, and let it be ℓ_i in the former case. Now, we construct C_{i+1} by removing ℓ_i' from C_i and taking the next literal of C, and we repeat. Note that since C has width at least $2w$, we find at least w such literals ℓ_i'. With probability at least $1-(1-\mu)^w \leq 1-\delta$, at least one of these literals is unmasked. The first such unmasked literal is as needed. ■

The final case to consider is when there is a subclause of width at most $\frac{1}{\mu\beta}\ln\frac{1}{\delta}$ (noting $\beta \leq 1/2$) for which there is a literal such that its negation is implied by the negation of the clause. In this case, as we elaborate on next, there is a small clause that we can learn that can be used to reduce the width of the first clause. If the clause is sufficiently wide and yet not witnessed to evaluate to true, it must be because there are many such literals. Using these learned clauses to eliminate these literals from the given clause enables us to find an equivalent clause that satisfies our width bound.

Analysis of Algorithm 1. We first note that the learned clauses are highly valid under D and contain clauses corresponding to $\ell_1 \wedge \cdots \wedge \ell_{k-1} \Rightarrow \ell_k$ whenever $\ell_1 \wedge \cdots \wedge \ell_{k-1} \rightsquigarrow \ell_k$ for $k \leq w$:

Lemma 18 *Suppose D is a distribution with a width-w $(\beta, 1-\frac{\gamma}{4(2n+1)^w})$ correlation gap. Let ψ be the conjunction of all clauses of width at most w that are witnessed to evaluate to false in at most a $\frac{\gamma\mu^w}{2(2n+1)^w}$-fraction of a sample of m_0 partial assignments from $M_\mu(D)$ (for m_0 as given in Algorithm 1). Then with probability at least $1-\delta/2$, ψ is $1-\gamma$-valid and contains all $1-\frac{\gamma}{4(2n+1)^w}$-valid clauses of width at most w, including specifically $C = C' \vee \ell$ such that $\neg C' \rightsquigarrow \ell$.*

Proof: Since each clause in ψ has width at most w, every literal in each such clause is simultaneously not set to $*$ by M_μ with probability at least μ^w; if a clause is not $1-\frac{\gamma}{(2n+1)^w}$ valid, then it is witnessed to evaluate to false on $M_\mu(D)$ with probability at least $\frac{\mu^w\gamma}{(2n+1)^w}$. Therefore, Hoeffding's inequality implies that it is only not witnessed to evaluate to false sufficiently often to eliminate it from ψ with probability at most $\frac{\delta}{4(2n+1)^w}$. Since there are fewer than $(2n+1)^w$ clauses of width at

function: `W-refute` (φ, w) decides whether or not there is a width-w refutation of CNF φ.
input : CNF φ, bound $p(n)$, $\epsilon, \delta, \gamma, \beta \in (0,1)$, partial assignments $\rho^{(1)}, \ldots, \rho^{(m_0+m_1)}$ from $M(D)$ for $m_0 = \frac{2w(2n+1)^{2w}}{\mu^{2w}\gamma^2} \ln \frac{4(2n+1)}{\delta}$ and $m_1 = \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$ where $w = \frac{1}{\mu\beta} \ln \frac{2m_1 \cdot p(n)}{\delta}$.
output : *Accept* or *Reject* (cf. Theorem 19)
begin
  Initialize ψ to an empty CNF.
  foreach *Clause C of width at most w* **do**
    FALSIFIED $\leftarrow 0$.
    for $i = 1, \ldots, m_0$ **do**
      if *C is falsified on $\rho^{(i)}$* **then**
        Increment *FALSIFIED*.
      end
    end
    if *FALSIFIED* $\leq \frac{\gamma\mu^w}{2(2n+1)^w} m_0$ **then**
      $\psi \leftarrow \psi \wedge C$.
    end
  end
  Initialize φ' to an empty CNF.
  foreach *Clause C from φ* **do**
    foreach *Clause C' from ψ* **do**
      if *$C' = C'' \vee \ell$ where $C'' \vee \neg\ell$ is a subclause of C* **then**
        $C \leftarrow C$ with $\neg\ell$ deleted.
      end
    end
    $\varphi' \leftarrow \varphi' \wedge C$.
  end
  FAILED $\leftarrow 0$.
  for $i = m_0 + 1, \ldots, m_0 + m_1$ **do**
    if `W-refute`$((\varphi' \wedge \psi)|_{\rho^{(i)}}, 2w)$ *rejects* **then**
      Increment *FAILED*.
      if *FAILED* $> \lfloor \epsilon \cdot m_1 \rfloor$ **then**
        return Reject
      end
    end
  end
  return Accept
end

Algorithm 1: Learn+RES

most w, by a union bound over these clauses, we find that with probability at least $1 - \delta/4$, these clauses are all $1 - \frac{\gamma}{(2n+1)^w}$-valid. Again, by a union bound over the clauses, this means that their conjunction, ψ, is $1 - \gamma$-valid.

Now, when a clause C of width at most w is $1 - \frac{\gamma}{4(2n+1)^w}$-valid (including $C = C' \vee \ell$ such that $\neg C' \rightsquigarrow \ell$), Hoeffding's inequality similarly guarantees that C will be witnessed to evaluate to false on $M_\mu(D)$ in a $\frac{\gamma\mu^w}{2(2n+1)^w}$-fraction of m_0 partial assignments with probability at most $\frac{\delta}{4(2n+1)^w}$.

Therefore, all of these clauses appear in ψ except with probability $\delta/4$, and a final union bound gives that both conditions hold with probability $1-\delta/2$. ■

The analysis of the algorithm is now a generalization of the analysis of the width-based algorithm discussed in Section 3.1. The main twist is that we may need to modify the resolution derivation by introducing learned clauses in order to obtain a low-width derivation.

Theorem 19 *Given an input CNF φ, a bound $p(n)$, $\mu, \epsilon, \gamma, \delta, \beta \in (0,1)$, and access to examples from $M_\mu(D)$ for a distribution D that has a width-$w = \frac{1}{\mu\beta}\ln\frac{2m_1 \cdot p(n)}{\delta}$ $\left(\beta, 1-\frac{\gamma}{4n(2n+1)^w}\right)$ correlation gap, (where $m_1 = \frac{1}{2\gamma^2}\ln\frac{2}{\delta}$) and given that either*

- *φ is satisfied by D with probability at least $\epsilon + 2\gamma$ or*
- *there exist CNFs ψ_0 and ψ_1 such that*
  - *ψ_0 consists of $1-\frac{\gamma}{4n(2n+1)^w}$-valid clauses and*
  - *ψ_1 is witnessed to evaluate to true with probability $1-\epsilon+2\gamma$ under $M_\mu(D)$*

  and there is a resolution refutation of size $p(n)$ of $\psi_0 \wedge \psi_1 \wedge \varphi$

Algorithm 1 decides which case holds with probability $1-\delta$, and runs in time $n^{O(\frac{1}{\mu\beta}\log\frac{p(n)}{\gamma\delta})}$.

Proof: We note that the running time bound is essentially immediate from the description of the algorithm, as the width-based algorithm, for width w', runs in time $O(n^{2w'})$. We therefore turn to considering the correctness of the algorithm. The first case is similarly simple: by Proposition 3, if φ is $\epsilon+2\gamma$-valid under D, then since the learned CNF ψ is $1-\gamma$-valid with probability at least $1-\delta/2$ by Lemma 18, then $\varphi\wedge\psi$ is $\epsilon+\gamma$-valid—meaning that with probability at least $\epsilon+\gamma$, a masked example ρ is drawn from $M_\mu(D)$ for which $(\varphi\wedge\psi)|_\rho$ is satisfiable. It then follows by Hoeffding's inequality that with probability greater than $1-\delta/2$, for at least an ϵ-fraction of the actual m_1 examples, $(\varphi\wedge\psi)|_\rho$ will be satisfiable, and therefore by the soundness of resolution, at least an ϵ-fraction of the iterations must fail to find a refutation (of the consequence $(\varphi'\wedge\psi)|_\rho$), leading the algorithm to reject as needed with probability at least $1-\delta$.

It only remains to establish that the algorithm accepts with probability at least $1-\delta$ when there exists a size-$p(n)$ refutation from some almost perfectly valid ψ_0 and some ψ_1 that is witnessed to evaluate to true with probability $1-\epsilon+2\gamma$ under $M_\mu(D)$. We assume (WLOG) that every clause of ψ_1 appears in the refutation. We now begin by writing the size-$p(n)$ proof Π as two subsets of clauses, Π_1 and Π_0 where Π_1 consists of clauses which either

1. contain a subclause C of width $w/2$ (for, recall, $w=\frac{1}{\mu\beta}\ln\frac{2m_1p(n)}{\delta}$) for which every further subclause $\ell\vee C'$ of C has ℓ balanced for $\neg C'$ or
2. contain a subclause of width w for which every (sub-)subclause C of width $w/2$ has a further subclause $C'\vee\ell$ for C' and ℓ satisfying $\neg C' \rightsquigarrow \ell$,

and Π_0 consists of the rest of the clauses.

We first note that the clauses in Π_1 (and in ψ_1) are simultaneously witnessed in most of the partial examples: we first note that by Hoeffding's inequality, with probability at least $1-\delta/2p(n)$, each clause in ψ_1 is witnessed to evaluate to true in at least $(1-\epsilon+\gamma)m_1$ partial examples; likewise, for the clauses of Π_1 which have subclauses for which every width $w/2$ subclause has an implied literal, by a union bound over all $2n$ literals, the probability that any of them is both a $(1-\frac{\gamma}{4n(2n+1)^w})$-implied literal indicated by Lemma 17 and false in any example is at most $\gamma/4$. (Note that with probability at least $1-\delta/(2p(n)m_1)$, every such clause has some such unmasked literal in each example.) Therefore, by Hoeffding's inequality, these clauses are all witnessed to evaluate to true in at least $(1-\gamma)m_1$ of the examples with probability at least $1-\delta/(2p(n)m_1)$.

Similarly, by Lemma 15, the clauses with β-balanced subclauses of width at least w are witnessed to evaluate to true in each partial example with probability at least $1-\delta/(2p(n)m_1)$. Hence by a union bound over the clauses and examples, every clause in Π_1 (and ψ_1) is ultimately (simultaneously) witnessed to evaluate to true in at least $(1-\epsilon)m_1$ of the examples considered in the main loop of the algorithm with probability $1-\delta/2$.

We now note that for every clause in Π_0, every subclause of width greater than w must have a subclause C of width $w/2$ with some literal ℓ such that for a further subclause $C' \vee \ell$, $\neg C' \rightsquigarrow \neg\ell$—every such C must have a subclause with a literal ℓ that is not balanced (or else the clause would be in Π_1 by the first condition) and the further subclause cannot have $\neg C' \rightsquigarrow \ell$ for every width $w/2$ subclause C (or else it would be in Π_1 by the second condition). Therefore, by Lemma 18, the clause $C' \vee \neg\ell$ is added to ψ by the algorithm. Moreover, since the clauses of ψ_0 are $1-\frac{\gamma}{4n(2n+1)^w}$-valid, and would reach width w after eliminating fewer than n such literals, for the clauses of ψ_0 in Π_0 there is a $\leq w$-CNF ψ_0' consisting of subclauses of every clause of ψ_0 appearing in Π_0 that (by Proposition 3) are $1-\frac{\gamma}{4(2n+1)^w}$-valid, and hence also added to ψ by Lemma 18.

Similarly, for every input clause C (from φ) in Π_0, we can always derive a subclause of width at most w using these clauses—as long as our subclause of C has width greater than w, some subclause of width $w/2$ must contain a subclause $C' \vee \ell$ for which a clause $C' \vee \neg\ell$ is in ψ, which we can use to reduce its size further, so the clauses of φ' corresponding to clauses of φ in Π_0 have width at most w.

To conclude, we observe that since all of the clauses of Π_1 (and ψ_1) are witnessed to evaluate to true in at least $(1-\epsilon)m_1$ of the partial examples considered by the algorithm in the main loop with probability $1-\delta/2$, for each such partial example ρ, by Proposition 14 $\Pi|_\rho$ is a resolution refutation of (restrictions of) the input and clauses of ψ_0, consisting only of (restrictions of) clauses from Π_0. Now, we know that there is a $\leq w$-CNF ψ_0' consisting of subclauses of every clause of ψ_0 that is included in ψ with probability at least $1-\delta/2$, and for each input clause in Π_0 there is a subclause C of width less than w that can be derived by the algorithm. Therefore, by induction on the steps of $\Pi|_\rho$, we know that by carrying out the steps of the proof with the derived subclauses, we can always derive some subclause C_i' of each ith clause in $\Pi|_\rho$, which is in Π_0, and hence has some further subclause of width at most w which can be derived in width $2w$: If the next step of $\Pi|_\rho$ is obtained by applying the cut rule to C_i and C_j, we can apply the cut rule to the derived width-w subclauses C_i' and C_j' to obtain a subclause of the next step of width at most $2w$. Since this next step is in Π_0, this clause can be further reduced to width w as noted above, completing the induction step. Since, finally, the only subclause of the empty clause (derived in the final step) is the empty clause itself, this ultimately yields a width-$2w$ refutation of $\Pi|_\rho$. Since the algorithm therefore derives the empty clause in at least $(1-\epsilon)m_1$ out of the m_1 examples with probability at least $1-\delta$, it therefore accepts with probability at least $1-\delta$ in this case, as needed. ■

Remarks on the algorithm. One could divide the derivation steps of our algorithm into "steps of $\Pi|_\rho$" and "intermediate derivations," where the intermediate derivations only involve the width w CNF ψ, and serve to find a width-w subclause of the next step of $\Pi|_\rho$. Based on this observation, we could have obtained a slightly more complicated algorithm that only uses dynamic programming over width-w clauses, that checks after each derivation to see if the clause can be reduced to width w by applications of the cut rule to clauses from ψ that strictly reduce the size of the intermediate clause (and discarding the clause if such a reduction is not possible). One might prefer it because its complexity is $O(n^{3w})$ rather than $O(n^{4w})$, as acheived by our algorithm.

In contrast to the simple algorithm of Section 3.1 (and more generally, the generic algorithms from prior work [15]) Theorem 19 also permits the proof to utilize an arbitrary *highly valid* CNF, in addition to one that is *witnessed* with probability $1 - \epsilon$. This is accomplished partially due to the simplicity of the distribution – we know that the clauses of any such CNF are either witnessed or are essentially equivalent to a short clause – and partially due to the simplicity of the masking process, which allows us to identify all such short clauses.

Clearly, our algorithm avoids the sources of intractability indicated by the various hardness results for the tasks in the two stages. It is easiest to see how the problem of Tseitin's hard tautologies [24] are avoided in the problem we consider: we only certify the validity of a DNF when there exists a small resolution proof for us to discover, and the proof complexity results merely establish that no such small resolution proof exists. So, such an example is "out of bounds" for our algorithm. The way we circumvent difficulties in learning is perhaps more interesting. Recall that Valiant's original work in this area [26] concerned algorithms that learned a collection of formulas and subsequently reasoned about what these learned formulas entailed. Following our previous work [15], our objective in learning is ultimately to determine whether or not *there exist* formulas that would suffice to complete a proof of a query. Here, we avoid a first common source of intractability by not aiming to find CNF representations of our hypotheses—essentially as done by "improper" learning algorithms. Arguably, this is the first step in circumventing the hardness of learning parities identified by Michael [20] (but note also that it is still not clear how to test whether a parity constraint of moderate size is satisfied in our affine distributions). Ultimately, the hardness of learning parities in the wRFA model [5] is circumvented by our structural result about the resolution proofs under our distributions: since we can show it suffices to only consider logarithmic-width clauses, we effectively only need to learn logarithmic-size parities, which is feasible in our partial-information model.

4 Deciding near-perfect validity of CNF queries

We now briefly note that our techniques also allow us to strengthen the results of Khardon and Roth [18] on learning to reason for the special case of distributions with a width-$O(\log n)$ $(\beta, 1 - 1/q(n))$ correlation gap (such as affine distributions) for a quasipolynomial function $q(n)$ and independent masking: we show how to decide whether a (general) CNF query is almost perfectly valid or significantly invalid. In particular, this avoids the requirement that the queries have small proofs, but the reader should notice that it applies only to establishing the validity of CNF (not DNF) queries, only for "nearly perfect" implications, and only for distinguishing "nearly perfect" queries from defective ones.

Theorem 20 *Given an input CNF φ, $\mu, \gamma, \beta \in [0,1]$, and access to examples from $M_\mu(D)$ for a distribution D with a width-$w = \frac{2}{\beta} \ln \frac{4|\varphi|}{\gamma}$ $(\beta, 1 - \frac{\gamma \mu^w}{8n(2n+1)^w})$ correlation gap, and given that either φ is at least $(1 - \frac{\gamma\mu^w}{8})$-valid under D or φ is at most $(1-\gamma)$-valid, Algorithm 2 decides which case holds with probability $1 - \delta$ in quasipolynomial time.*

Proof: We begin by noting that the running time of the algorithm is indeed quasipolynomial in $|\varphi|$, $1/\beta$, $1/\gamma$, $1/\mu$, and $\log 1/\delta$. Now, towards correctness, we observe that by Lemma 18, the formula ψ consisting of the conjunction of learned clauses is $1 - 3\mu^w\gamma/8$ valid with probability $1 - \delta/2$. We further note that $\varphi \wedge \psi \models \varphi'$ since the clauses of φ' have a simple resolution derivation from clauses of φ and ψ.

input : CNF φ, $\epsilon, \delta, \gamma, \beta \in (0,1)$, list of partial assignments $\rho^{(1)}, \ldots, \rho^{(m_0+m_1)}$ from $M(D)$ for $m_0 = \frac{32w(2n+1)^{2w}}{\mu^{4w}\gamma^2} \ln \frac{4n+2}{\delta}$ and $m_1 = \frac{32}{(\mu^w\gamma)^2} \ln \frac{2}{\delta}$ where $w = \frac{2}{\beta} \ln \frac{4|\varphi|}{\gamma}$ and $|\varphi|$ denotes the number of clauses in φ.

output: *Accept* if φ is at least $(1 - \frac{\gamma\mu^w}{8})$-valid under D or *Reject* φ is at most $(1-\gamma)$-valid

begin

Initialize ψ to an empty CNF.

foreach *Clause C of width at most w* **do**

FALSIFIED $\leftarrow 0$.

for $i = 1, \ldots, m_0$ **do**

if *C is falsified on $\rho^{(i)}$* **then**

Increment *FALSIFIED*.

end

end

if *FALSIFIED* $\leq \frac{\gamma\mu^{2w}}{4(2n+1)^w} m_0$ **then**

$\psi \leftarrow \psi \wedge C$.

end

end

Initialize φ' to an empty CNF.

foreach *Clause C from φ* **do**

foreach *Clause C' from ψ* **do**

if *$C' = C'' \vee \ell$ where $C'' \vee \neg\ell$ is a subclause of C* **then**

$C \leftarrow C$ with $\neg\ell$ deleted.

end

end

$\varphi' \leftarrow \varphi' \wedge C$.

end

FALSIFIED $\leftarrow 0$.

for $i = m_0 + 1, \ldots, m_0 + m_1$ **do**

if *Any clause of φ' is falsified on $\rho^{(i)}$* **then**

Increment *FALSIFIED*.

if *FALSIFIED* $> (5\mu^w\gamma/8)m_1$ **then**

return Reject

end

end

end

return Accept

end

Algorithm 2: CNF-Eval

Let us consider the first case, when φ is at least $(1 - \mu^w\gamma/8)$-valid. Here, since $\varphi \wedge \psi \models \varphi'$, φ' is at least $(1 - \mu^w\gamma/2)$-valid by Proposition 3. Therefore, by Hoeffding's inequality, in a sample of size m_1 from D, it is satisfied in at least $(1 - 5\mu^w\gamma/8)m_1$ of the (underlying, full) examples with probability at least $1 - \delta/2$. Since the partial examples are consistent with these full examples, this means in particular that no clause of φ' can be witnessed to evaluate to false in more than $(5\mu^w\gamma/8)m_1$ of the partial examples, so we see that the algorithm accepts with probability at least

$1-\delta$ in this case, as needed.

Now, turning to the second case, we split the clauses of φ into two formulas, φ_0 and φ_1 as follows: let φ_1 consist of those clauses having either $w/2$ literals that are balanced given any subset of the rest (of the $w/2$) under D or a subclause of width w such that every width $w/2$ subclause has a literal ℓ such that for a further subclause $C' \vee \ell$, $\neg C' \rightsquigarrow \ell$, and let φ_0 consist of the rest. Clauses in φ_1 are simultaneously true in most of the partial examples: essentially by Lemma 15 with $\mu = 1$ and the definition of implied literals, each clause in φ_1 is true in each underlying example (though not necessarily witnessed) with probability at least $1 - \gamma/(4|\varphi|)$. Hence by a union bound over the clauses, every clause in φ_1 is simultaneously true with probability $1 - \gamma/4$. By contrast, though, since φ is at most $1-\gamma$-valid, we see that with probability at least $3\gamma/4$, some clause in φ_0 must evaluate to false.

Let φ_0' be the clauses of φ' corresponding to clauses from φ_0. For every subclause C of a clause of φ_0 of width greater than w, some further subclause $C' \vee \ell$ of width at most $w/2$ must have $\neg C' \rightsquigarrow \neg\ell$, as if every literal is balanced in any small subclause, the orignal clause would have been in φ_1 by the first condition, and if some width w subclause C had $\neg C' \rightsquigarrow \ell$ in some further subclause of every width $w/2$ subclause of C, it would be in φ_1 by the second condition. Much as in the proof of Theorem 19, by an analogue of Lemma 18, since this $C' \vee \neg\ell$ is $(1 - \frac{\gamma\mu^w}{8(2n+1)^w})$-valid, it is in ψ with probability $1-\delta/2$, and hence there is some clause of ψ that can be used to eliminate one of the literals on this width-$w/2$ subclause. Therefore, we find that the clauses of φ_0' must have width at most w with probability $1-\delta/2$.

Now, since M_μ reveals each literal with probability μ independently, given that the clauses of φ_0' have width at most w, each clause in φ_0' is completely revealed with probability at least μ^w, and hence some clause in φ_0' is witnessed to evaluate to false in the partial examples sampled from $M_\mu(D)$ with probability at least $3\mu^w\gamma/4$. Therefore, by Hoeffding's inequality, some clause is witnessed to evaluate to false in at least a $(5\mu^w\gamma/8)$-fraction of the m_1 partial examples with probability at least $1-\delta/2$ given that the clauses have width at most w, which we argued occurs with probability at least $1-\delta/2$. Therefore, the algorithm rejects with probability at least $1-\delta$ in this case, as needed. ■

5 Directions for future research

Several problems present themselves as natural directions for future work. The most pressing of these is, *can the restriction to distributions with a correlation gap be lifted?* That is, how can we efficiently reason about "medium-strength" correlations? Although the ultimate objective of such work would be to strengthen these results to the distribution-free PAC setting, any work that handled a class of distributions that exhibited such correlations would also be of interest. A similar direction would be to obtain results for a more general class of masking processes; although it seems that our results generalize to masking distributions that simultaneously reveal any width-w set of literals with non-negligible probability (for $w = \Omega(\log n)$) such as w-wise independent distributions (Wigderson and Yehudayoff [27] make a similar observation about their algorithm), it would be desirable to find other, perhaps weaker properties that would also permit relatively efficient algorithms.

Of course, the results of this work beg the question so far as the classical (quasi)automatizability of resolution is concerned. Although there are families of counterexamples [8, 3] showing that a purely width (and/or treelike) based approach to finding small resolution proofs such as pursued

by Ben-Sasson and Wigderson [6] cannot beat the current best-known bound of $n^{O(\sqrt{n \log n})}$, it does not rule out other approaches. Since our algorithm and analysis essentially establish that every resolution proof over distributions with a correlation gap has a low-width approximate version using the learned clauses, it seems significant for our algorithm that the learned formula ψ may not have a small-width derivation. Unfortunately, it is not clear how one might hope to exploit this in the absence of a distribution. Still, if *any* algorithm could find resolution derivations in quasipolynomial time, then using the results of our previous work [15], this would also immediately resolve both of the questions suggested in the previous paragraph.

The other natural direction in which one might hope to strengthen our results involves extending them to stronger proof systems than resolution, such as cutting planes or polynomial calculus (with resolution, aka PCR). We already observed in previous work [15] that there are natural fragments of these proof systems (most already well studied) for which the combined learning and reasoning problem is tractable. The question would be whether, as with width-restricted resolution, we could use these algorithms as a starting point to obtain algorithms for the unrestricted proof system in the context of reasoning about a distribution.

There are, of course, limits to what we can hope for: the strong non-automatizability results for systems such as bounded-depth Frege based on cryptographic hardness assumptions [7, 9, 19] are established by formulas equipped with a natural background distribution (over the variables encoding parameters generated by the Diffie-Hellman key exchange protocol [12]), in which moreover the pattern of masking is independent of the underlying example. (The bits are *not* masked with the same probability as in M_μ, though.) Although these results were not stated in the context of PAC-Semantics and the distributions were ignored, they should carry over to an appropriate generalization of the setting we considered here, still within the framework of the prior work [15]. One might reasonably ask if some analogous negative result can be proved for M_μ based on leakage-resilient cryptography (as M_μ leaks a constant fraction of the secret parameters).

Acknowledgements

The author would like to thank Paul Beame, Eli Ben-Sasson, and Leslie Valiant for comments and conversations that helped shape this work.

References

[1] Michael Alekhnovich and Alexander A. Razborov. Resolution is not automatizable unless W[P] is tractable. *SIAM J. Comput.*, 38(4):1347–1363, 2008.

[2] Misha Alekhnovich, Mark Braverman, Vitaly Feldman, Adam R. Klivans, and Toniann Pitassi. The complexity of properly learning simple concept classes. *JCSS*, 74(1):16–34, 2008.

[3] Albert Atserias and María Luisa Bonet. On the automatizability of resolution and related propositional proof systems. *Inf. Comp.*, 189:182–201, 2004.

[4] Paul Beame, Henry Kautz, and Ashish Sabharwal. Towards understanding and harnessing the potential of clause learning. *JAIR*, 22:319–351, 2004.

[5] Shai Ben-David and Eli Dichterman. Learning with restricted focus of attention. *JCSS*, 56(3):277–298, 1998.

[6] Eli Ben-Sasson and Avi Wigderson. Short proofs are narrow – resolution made simple. *J. ACM*, 48(2):149–169, 2001.

[7] Maria Luisa Bonet, Carlos Domingo, Ricard Gavaldá, Alexis Maciel, and Toniann Pitassi. Non-automatizability of bounded-depth frege proofs. *Comput. Complex.*, 13:47–68, 2004.

[8] María Luisa Bonet and Nicola Galesi. Optimality of size-width tradeoffs for resolution. *Computational Complexity*, 10(4):261–276, 2001.

[9] Maria Luisa Bonet, Toniann Pitassi, and Ran Raz. On interpolation and automization for Frege proof systems. *SIAM J. Comput.*, 29(6):1939–1967, 2000.

[10] Matthew Clegg, Jeff Edmonds, and Russell Impagliazzo. Using the Gröbner basis algorithm to find proofs of unsatisfiability. In *Proc. 28th STOC*, pages 174–183, 1996.

[11] Scott E. Decatur and Rosario Gennaro. On learning from noisy and incomplete examples. In *Proc. 8th COLT*, pages 353–360, 1995.

[12] Whitfield Diffie and Martin Hellman. New directions in cryptography. *IEEE Trans. Inform. Theory*, 22:423–439, 1976.

[13] Zeev Dvir, Anup Rao, Avi Wigderson, and Amir Yehudayoff. Restriction access. In *Proc. 3rd ITCS*, 2012.

[14] Zvi Galil. On resolution with clauses of bounded size. *SIAM J. Comput.*, 6:444–459, 1977.

[15] Brendan Juba. Implicit learning of common sense for reasoning. To appear in IJCAI'13, 2013. Preliminary version: *Learning implicitly in reasoning in PAC-Semantics*, arXiv:1209.0056v1 [cs.AI].

[16] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Towards efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[17] Roni Khardon and Dan Roth. Learning to reason. *J. ACM*, 44(5):697–725, 1997.

[18] Roni Khardon and Dan Roth. Learning to reason with a restricted view. *Machine Learning*, 35:95–116, 1999.

[19] Jan Krajíček and Pavel Pudlák. Some consequences of cryptographical conjectures for S^1_2 and EF. In D. Leivant, editor, *Logic and Computational Complexity*, volume 960 of *LNCS*, pages 210–220. Springer, Berlin, 1995.

[20] Loizos Michael. Partial observability and learnability. *Artificial Intelligence*, 174(11):639–669, 2010.

[21] Ankur Moitra and Michael Saks. A polynomial time algorithm for lossy population recovery. arXiv:1302.1515, 2013.

[22] Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *JCSS*, 61(2):217–235, 2000.

[23] Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1–2):273–302, 1996.

[24] G. S. Tseitin. On the complexity of derivation in propositional calculus. In A. O. Slisenko, editor, *Studies in constructive mathematics and mathematical logic, part 2*, pages 115–125. Consultants Bureau, New York, 1970.

[25] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 18(11):1134–1142, 1984.

[26] Leslie G. Valiant. Robust logics. *Artificial Intelligence*, 117:231–253, 2000.

[27] Avi Wigderson and Amir Yehudayoff. Population recovery and partial identification. In *Proc. 53rd FOCS*, 2012.